Truncation, Information, and the Coefficient of Variation

M.J. Bayarri
M.H. DeGroot
P.K. Goel

Abstract The Fisher information in a random sample from the truncated version of a distribution that belongs to an exponential family is compared with the Fisher information in a random sample from the untruncated distribution. Conditions under which there is more information in the selection sample are given. Examples involving the normal and gamma distributions with various selection sets, and the zero-truncated binomial, Poisson, and negative binomial distributions are discussed. A property pertaining to the coefficient of variation of certain discrete distributions on the non-negative integers is introduced and shown to be satisfied by all binomial, Poisson, and negative binomial distributions.

1 Introduction

Consider the basic statistical problem in which a random variable $X$ is distributed over a certain population according to the (generalized) density $g(x \mid \theta)$ and it is desired to make inferences about the unknown value of the parameter $\theta$ which lies in the parameter space $\Omega$. In the usual statistical analysis, it is assumed that the observed data form a random sample from the density $g(x \mid \theta)$. In many situations, however, observations are obtained only from certain selected portions of the underlying population, either because experimental conditions make it impossible to obtain data from the entire population or because experimenters choose to restrict the observations in this way in their experimental design.

In this paper we will consider problems in which the observations are restricted to lie in a specified subset $S$ of the sample space of $X$. Let

$$s(\theta) = \Pr(X \in S \mid \theta) \quad \text{for} \quad \theta \in \Omega,$$

(1.1)
where it is assumed that $s(\theta) > 0$ throughout $\Omega$. The statistical analysis will then be based on a random sample $Y_1, \ldots, Y_n$ from the following truncated version of the density $g$:

$$
f(y \mid \theta) = \begin{cases} 
g(y \mid \theta)/s(\theta) & \text{for } y \in S, \\
0 & \text{otherwise.} \end{cases} \tag{1.2}
$$

The model (1.2) is called a truncation model or a selection model, and a random sample from (1.2) is called a selection sample. The set $S$ is called the selection set.

Selection samples occur frequently in practice. Several examples are given in Bayarri and DeGroot (1987a,c), together with a discussion of the Bayesian approach to their analysis.

Our central focus here is to compare the experiment in which a random sample $X_1, \ldots, X_n$ is drawn from the unrestricted density $g(x \mid \theta)$ with the experiment in which a selection sample $Y_1, \ldots, Y_n$ is drawn from the truncation model $f(y \mid \theta)$ in order to determine which of these two experiments is more informative about $\theta$. Many different concepts of the information in an experiment have been discussed in the statistical literature, including sufficiency and the comparison of experiments, as developed by Blackwell (1951, 1953); Kullback-Leibler information (Kullback, 1968); and the widely-used Fisher information. Some relationships among these concepts and further references can be found in Goel and DeGroot (1979). Here we will restrict our study to problems in which $\theta$ is a real-valued parameter and the comparison is based on the Fisher information in each experiment. The investigation of the information in selection samples was introduced in Bayarri and DeGroot (1987b), where comparisons based on both Fisher information and other concepts of information were carried out. Some of the results to be discussed here were mentioned there with the details omitted. In closely related subsequent work, Patil and Taillie (1987) calculate the Fisher information for a wide variety of weighted distributions, including some truncation models. Their paper, like this one, concentrates on exponential families of distributions.

This paper has two major purposes. The first is to illuminate the effects of truncation on the information in the most widely used distributions in statistical practice, including the normal, gamma, binomial, Poisson, and negative binomial. We shall accomplish this by studying the Fisher information in truncated exponential families. The second purpose is really a happy bonus. We will present a simple and fascinating property that is satisfied by all binomial, Poisson, and negative binomial distributions, but (relatively speaking) few other discrete distributions. This property, which we discovered in our study of the Fisher information in truncated versions of these standard discrete distributions with the zero class missing, is defined as follows: